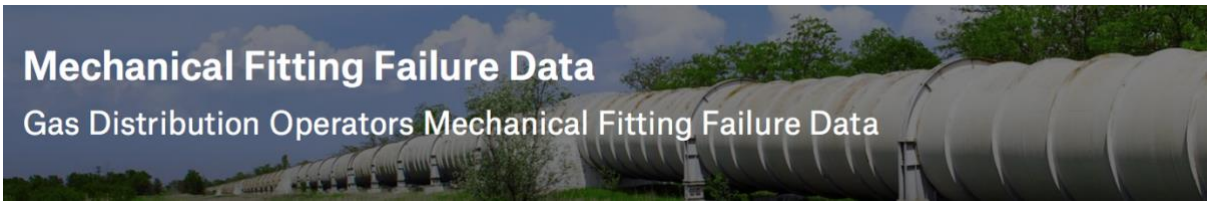
Project: Capstone Project 2: Milestone Report

Mechanical Fitting Failure Classification



Problem Statement: Code of Federal Regulations (49 CFR Parts 191, 192) requires gas distribution pipeline operators to submit reports on an annual basis of all hazardous leaks that involve a mechanical fitting (DOT Form PHMSA F-7100.1-2).

Our goal is to classify mechanical fitting failure in the gas pipeline so that we can identify in-advance reason for the leak, when it happened and may be how often since installation? If there any fittings that failed during the first year of operation, then that issue could be either design related or material related from the manufacturer, which should get reported to CFR.

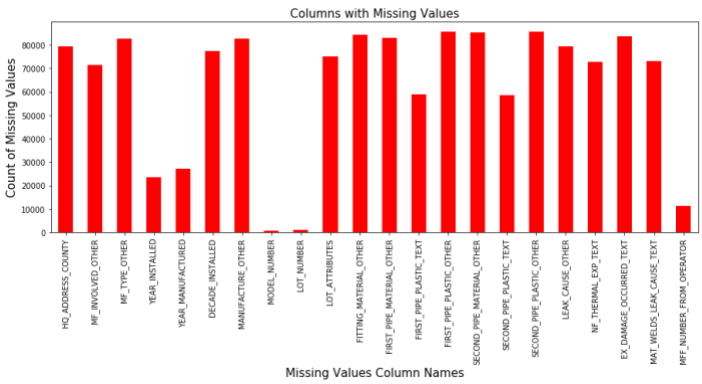
Benefit to Client: This problem is relevant to all pipeline operators who are responsible for transporting oil or gas to various locations using their pipeline network safely, reliably and long-term.

By training a model, if we could identify in advance which mechanical fittings leak frequently, what causes the leak to happen and after how many years of operation the leak usually occurs, then this will help our client to proactively manage repair/ maintenance schedule on specific fittings that are impacted. It will also allow client to engage manufacturers to research and develop improved fittings that offer better life and provide advance notification to pipeline operators before a fitting leaks.

Data Story:

This dataset was available on Kaggle in csv format. My mentor discovered this dataset and recommended me to work on it. This dataset contained 85,611 observations and 54 columns detailing various attributes about mechanical fittings used in the gas pipeline, some as old as 165 years old.

One of the major challenges with this dataset was that it was entirely in text format. As an example, failure date when the leak occurred, installation date when the fittings were installed and pipe nominal sizes, all were presented in text format which had to be transformed into date time format and numerical values.



Moreover, there were multiple columns where information regarding leak and the manufacturers were captured in two separate columns. One represented by column label ending with _TEXT while the other as a subset represented by _OTHER. The columns ending with _TEXT had multiple missing values which had to be replaced by values in _OTHER columns.

To get a better understanding of the missing values, I wrote a function missing_dashboard which gave an overview on missing columns and unique values as shown in the table below.

The size of this table is dependent on a user defined variable 'missing_percentage' which can be selected anywhere between 0 to 1. For this table as an example, I chose 0.7 as the missing_percentage, which collected information on 15 missing columns out of 22.

```
Average (base) Missing Percentage: 0.73
Number of Missing Columns
(Random with Missing Percentage): 15
```

	missing_values	missing_percentage	unique_values_in_missing_columns	available_values_in_missing column
HQ_ADDRESS_COUNTY	79345	0.93	84	6266
MF_INVOLVED_OTHER	71434	0.83	741	14177
MF_TYPE_OTHER	82557	0.96	328	3054
DECADE_INSTALLED	77247	0.90	10	8364
MANUFACTURE_OTHER	82683	0.97	217	2928
LOT_ATTRIBUTES	75017	0.88	738	10594
FITTING_MATERIAL_OTHER	84411	0.99	58	1200
FIRST_PIPE_MATERIAL_OTHER	82845	0.97	45	2766
FIRST_PIPE_PLASTIC_OTHER	85557	1.00	31	54
SECOND_PIPE_MATERIAL_OTHER	85274	1.00	56	337
SECOND_PIPE_PLASTIC_OTHER	85568	1.00	24	43
LEAK_CAUSE_OTHER	79267	0.93	822	6344
NF_THERMAL_EXP_TEXT	72766	0.85	2	12845
EX_DAMAGE_OCCURRED_TEXT	83503	0.98	2	2108
MAT_WELDS_LEAK_CAUSE_TEXT	73060	0.85	2	12551

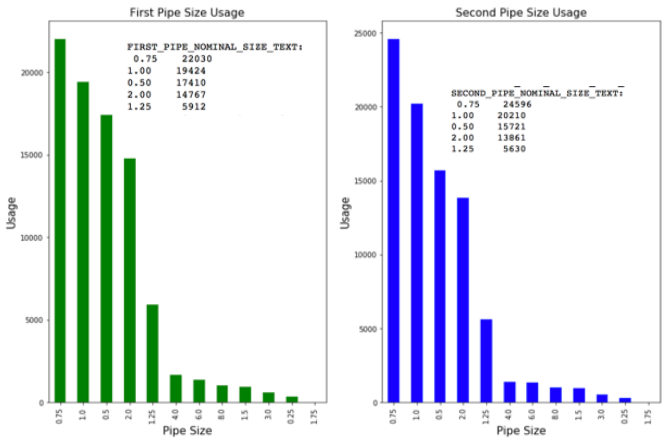
As we can see above, there are multiple unique values in the missing columns which can be used to replace missing values in other columns. To find this out, I wrote a function to help identify missing replacement values from _OTHER columns to fill missing values in the _TEXT column.

1. missing_value_replacement
2. missing_value_index_from_column_1
3. getvalues_from_column_2_using_missing_index_from_column_1
4. missing_replacement_value_and_index_column_2

As a result, we were able to replace a total of **18,789** missing values in a total of 9 columns.

Exploratory Analysis:

After replacing some missing values, my focus shifted on wrangling dates and getting insights from the data. I was interested in exploring which states in the US had maximum number of leaks by manufacturers. Which Manufacturers accounted for maximum number of leaks and what was the primary cause for the leak. Below, are some graphs which will help us answer these questions.

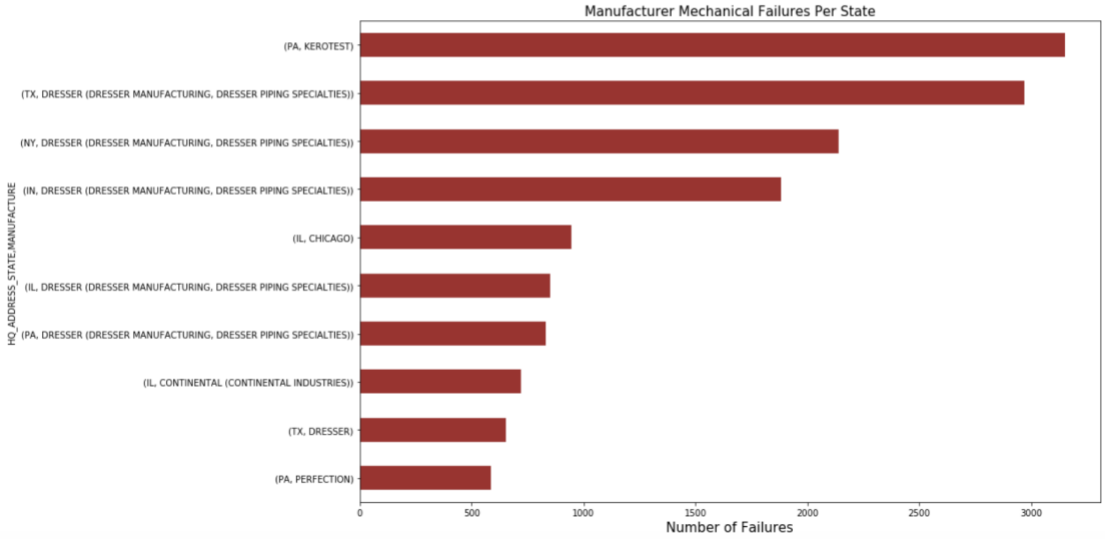


Most Used Pipe Size:

After converting pipe size from text format to numerical value, pipe size 0.75 showed most usage among pipeline operators.

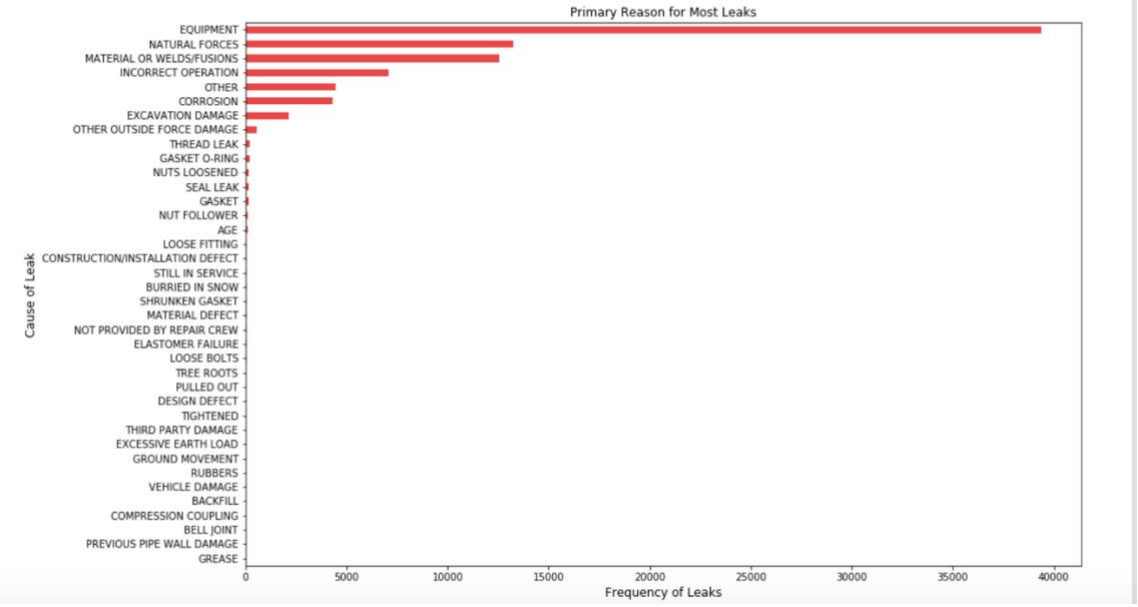
The top 10 known manufacturers by state where leak occurred:

```
HQ_ADDRESS_STATE  MANUFACTURE
PA                KEROTEST                                                  3152
TX                DRESSER (DRESSER MANUFACTURING, DRESSER PIPING SPECIALTIES)  2972
NY                DRESSER (DRESSER MANUFACTURING, DRESSER PIPING SPECIALTIES)  2138
IN                DRESSER (DRESSER MANUFACTURING, DRESSER PIPING SPECIALTIES)  1881
IL                CHICAGO                                                    944
Name: MANUFACTURE, dtype: int64
```



Top 15 reasons for the leak:

```
****Top 15 reasons for leak****
                          Reason_Count  % Reason_Count
EQUIPMENT                        39370       45.987081
NATURAL FORCES                   13230       15.453622
MATERIAL OR WELDS/FUSIONS        12551       14.660499
INCORRECT OPERATION               7070        8.258285
OTHER                             4464        5.214283
CORROSION                         4330        5.057761
EXCAVATION DAMAGE                 2123        2.479822
OTHER OUTSIDE FORCE DAMAGE         575        0.671643
THREAD LEAK                        225        0.262817
GASKET O-RING                      223        0.260481
NUTS LOOSENED                      180        0.210253
SEAL LEAK                          163        0.190396
GASKET                             144        0.168203
NUT FOLLOWER                       133        0.155354
AGE                                 90        0.105127
```

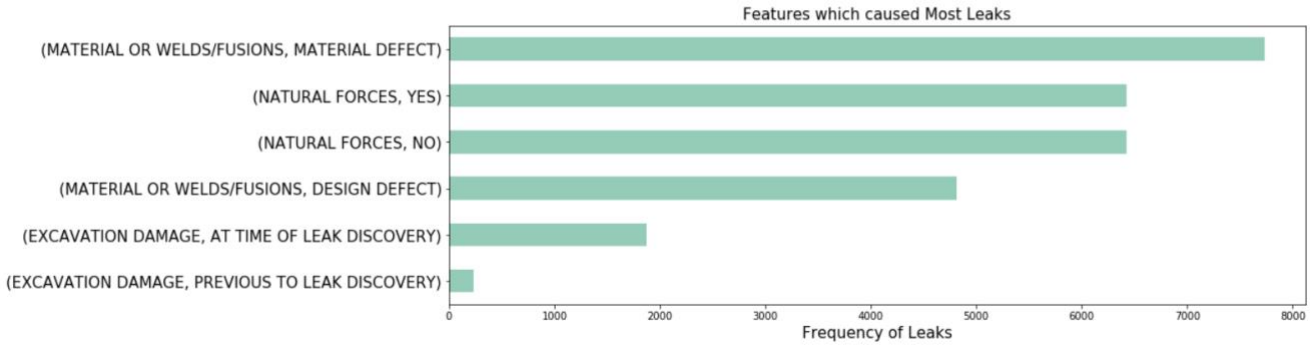


Features which caused leak in the pipeline:

****Features which caused leak****

LEAK_CAUSE_TEXT	ADDITIONAL_LEAK_FEATURES	Feature_Count
MATERIAL OR WELDS/FUSIONS	MATERIAL DEFECT	7734
NATURAL FORCES	YES	6424
	NO	6421
MATERIAL OR WELDS/FUSIONS	DESIGN DEFECT	4817
EXCAVATION DAMAGE	AT TIME OF LEAK DISCOVERY	1872
	PREVIOUS TO LEAK DISCOVERY	236

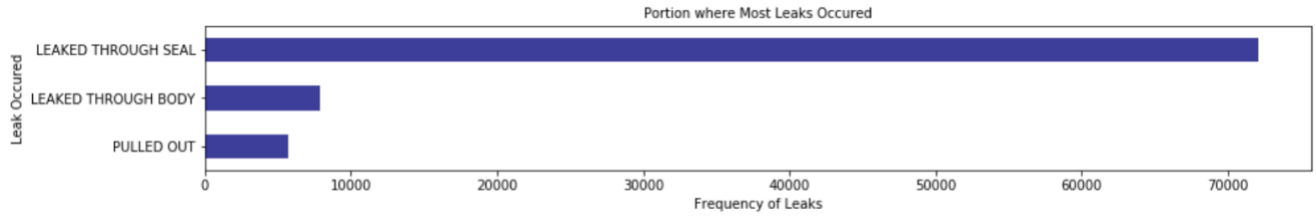
LEAK_CAUSE_TEXT	ADDITIONAL_LEAK_FEATURES	%_Feature_Count
MATERIAL OR WELDS/FUSIONS	MATERIAL DEFECT	9.033886
NATURAL FORCES	YES	7.503709
	NO	7.500204
MATERIAL OR WELDS/FUSIONS	DESIGN DEFECT	5.626613
EXCAVATION DAMAGE	AT TIME OF LEAK DISCOVERY	2.186635
	PREVIOUS TO LEAK DISCOVERY	0.275666



Portion where most leaks occurred:

****Portion where leak occured****

	Occurred_Count	% Occurred_Count
LEAKED THROUGH SEAL	72062	84.173763
LEAKED THROUGH BODY	7867	9.189240
PULLED OUT	5682	6.636998



Observation:

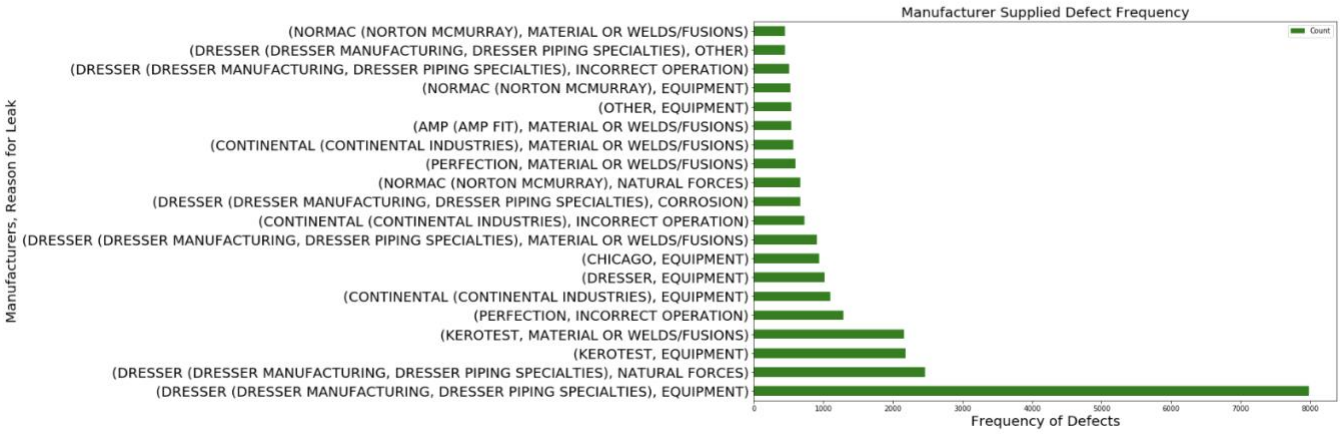
Maximum Impact:

1. 39.4k or 46% leaks were caused due to **equipment failure**, where features of failure are not known.
2. 7.7k or 9% leak were caused due to **welding defects**
3. 72.1k or 84% **leaked through the seal**.

Manufacturer Defects:

Manufacturer Supplied Defect Frequency:

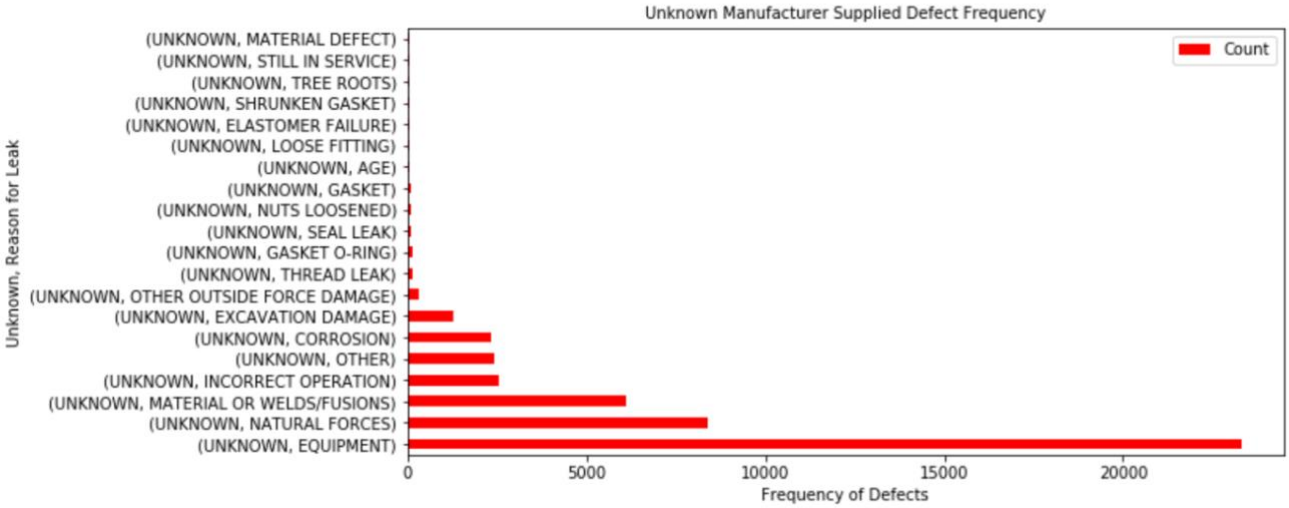
MANUFACTURE	LEAK_CAUSE_TEXT	Count
DRESSER (DRESSER MANUFACTURING, DRESSER PIPING ...	EQUIPMENT	7985
	NATURAL FORCES	2461
KEROTEST	EQUIPMENT	2177
	MATERIAL OR WELDS/FUSIONS	2162
PERFECTION	INCORRECT OPERATION	1285



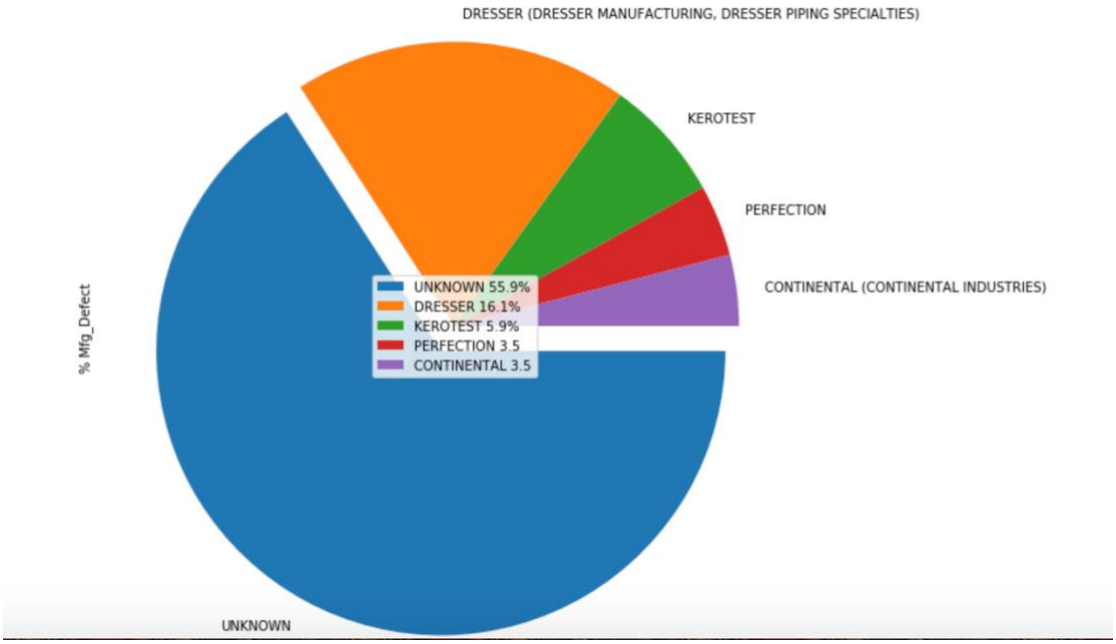
Unknown Manufacturer Defects:

Unknown Manufacturer Supplied Defect Frequency:

MANUFACTURE	LEAK_CAUSE_TEXT	Count
UNKNOWN	EQUIPMENT	23328
	NATURAL FORCES	8408
	MATERIAL OR WELDS/FUSIONS	6108
	INCORRECT OPERATION	2537
	OTHER	2434



Percentage of Manufacturer Defects:

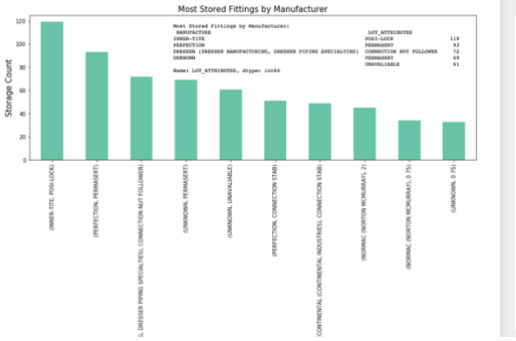


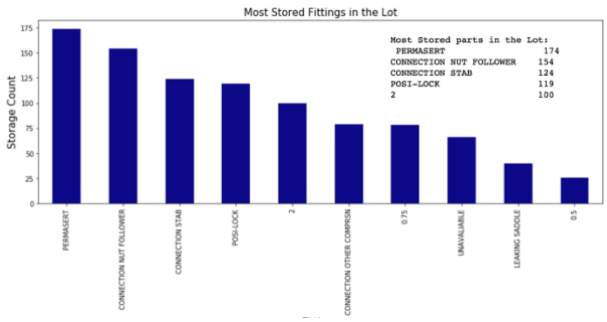
Observation:

1. 56% defects were caused by Unknown manufacturers, majority related to equipment failure.
2. Dresser and Kerotest accounted for 16% and 6% defects related to (equipment, natural forces) and (equipment, welding) respectively.

Most Stored Fittings by Manufacturer and By Lot:

After cleaning lot attributes and exploring manufacturers, Permasert was observed to be the most stored fitting in the lot supplied by two primary manufacturers Dresser and Perfection.
While, per manufacturer individually inter-tite, the manufacturer of POSI-LOCK carried maximum fittings in the lot compared to other manufacturers.





Relationship between dates (Installation Vs Report Vs Filing):

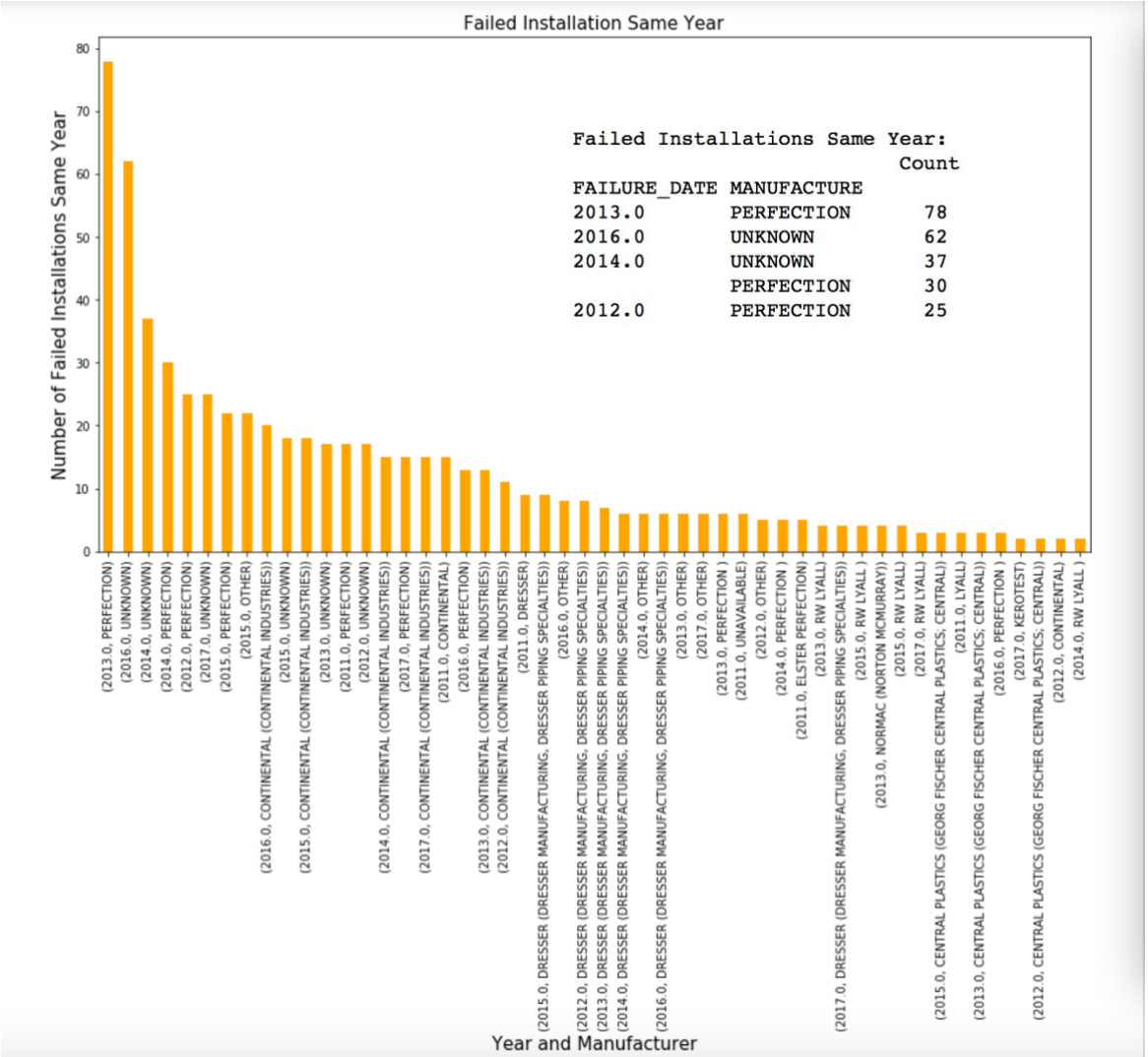
It was observed that there was a significant delay (12 years +) between when the failure occurred Vs when it was reported and filed.
After further investigation, it was clear that majority of the reports were filed the same day as the failure date. Hence, filing date was corrected and matched with the failure date to avoid any date lag in our observation.

Oldest Installation:

First Mechanical fitting was installed 165 years ago in the year 1851. There were 6 installations done back then by an Unknown manufacturer.

Failed Installation Same Year:

Below graph shows failed installation by the manufacturer in the same year of installation.



Observation:

1. Perfection had maximum number of failed installations in 2013.
2. Dresser and Continental and Perfection had multiple failed installations between 2011 and 2017 as shown in the graph above.

Additionally , 20 columns were dropped after extracting missing values and 4 new columns were added to drive meaningful insights from the raw data.

----------End of Report----------